THE MISSING HORIZON IN RANDOM RESHUFFLING

Pahan Dewasurendra
Johns Hopkins University

ABSTRACT

Random reshuffling now has an upper bound that dominates stochastic gradient descent at every constant step $\eta \le 1/(6L)$ and every finite horizon. Its matching lower bound was left open. We solve this problem for the exact all-inner average used by the upper bound. The answer is not the published horizon-independent stochastic envelope. If $\alpha = \eta n L \le 1$, the fixed-step minimax stochastic error is

$$\Theta\big(\eta^2 n L \sigma_\star^2 \left[K^{-1} + \min\{(\alpha K)^2, 1\}\right]\big) .$$

It decreases, increases, and then plateaus across three horizon regimes. Its minimum occurs at $K \asymp \alpha^{-2/3}$, before the usual mixing scale α^{-1}. For $\alpha \ge 1$, the law is $\Theta(\eta\sigma_\star^2)$. The upper bound follows from an exact finite-population bridge identity and a dimension-free control of all state memory as a Lipschitz remainder. The lower bound combines a centered common-Hessian bridge with a rectifying piecewise quadratic in two dimensions. An exact telescoping identity explains why common-Hessian quadratics alone miss the eventual floor. Together with the deterministic term, this gives a complete minimax characterization for all horizons and this full step range. Optimizing the step recovers the best known tuned rate up to the unavoidable initial-error cap, while the fixed-step theorem reveals a nonmonotone transient hidden by that rate.

1 INTRODUCTION

Random reshuffling processes every component of a finite sum exactly once per epoch in a fresh random order. It is the common practical implementation of stochastic gradient descent, and it is often faster than sampling with replacement. A long line of work has sought to explain this advantage (Gürbüzbalaban et al., 2021; Nagaraj et al., 2019; Haochen & Sra, 2019; Mishchenko et al., 2020; Nguyen et al., 2021; Safran & Shamir, 2020; 2021).

The recent result of Liu (2026) resolves the upper-bound side for smooth convex finite sums. For any $\eta \le 1/(6L)$ and any number of epochs K, it bounds the exact all-inner average by

$$\frac{D^2}{\eta n K} + \min\{\eta, \eta^2 n L\}\sigma_\star^2. \tag{1}$$

This is never slower than the corresponding bound for with-replacement SGD. The same paper identifies the matching lower characterization for all steps and horizons as an important open problem. Existing lower bounds cover small steps, sufficiently many epochs, and averages of epoch-boundary iterates (Cha et al., 2023). They do not characterize the all-inner output in (1).

A natural conjecture is that (1) is minimax-sharp for each fixed (η, K). It is false. All-inner averaging sees a fresh without-replacement bridge in every epoch, but the epoch-end state remembers earlier bridges. Before curvature has time to rectify these fluctuations, the two effects cancel more strongly than the envelope records. The correct stochastic law for $\alpha = \eta n L \le 1$ is

$$\eta^2 n L \sigma_\star^2 \left[\frac{1}{K} + \min\{(\alpha K)^2, 1\}\right]. \tag{2}$$

It has three phases:

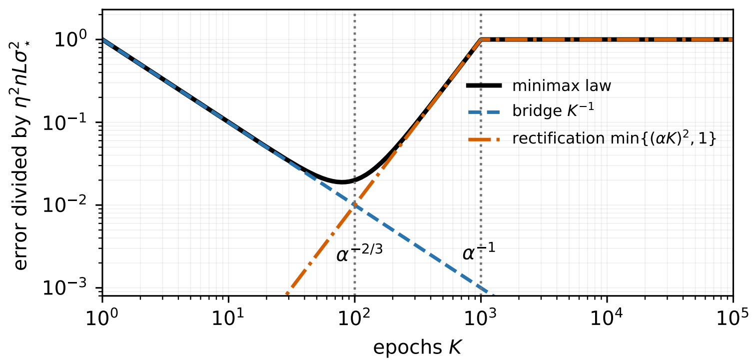

Figure 1: The normalized stochastic law for $\alpha = 10^{-3}$. The fresh bridge decays as K^{-1}, while rectification grows quadratically and then saturates. Their sum reaches its minimum at $\alpha^{-2/3}$, before the mixing scale α^{-1}.

horizon	dominant mechanism	stochastic error
$K \lesssim \alpha^{-2/3}$	fresh bridge	$\eta^2 nL\sigma_\star^2/K$
$\alpha^{-2/3} \lesssim K \lesssim \alpha^{-1}$	rectification	$\eta^2 nL\sigma_\star^2(\alpha K)^2$
$K \gtrsim \alpha^{-1}$	rectified floor	$\eta^2 nL\sigma_\star^2$

Thus more epochs can temporarily make the fixed-step minimax error worse. The minimum is smaller than the eventual floor by $\alpha^{2/3}$.

Our proof isolates the two mechanisms. For the upper bound, freeze every component gradient at a minimizer. The resulting permutation prefixes reset to zero after each epoch and have an exact bridge-area variance of $\eta^2(n+1)\sigma_\star^2/12$. Every nonlinear and state-memory effect is a single Lipschitz remainder. A prefix-sum operator controls this remainder when αK is small, and the general bound of Liu (2026) covers the remaining horizons.

For the lower bound, a centered common-Hessian quadratic generates the $1/K$ bridge term. It also satisfies an exact identity that makes its all-inner average proportional to its final state, so its error eventually decays rather than reaching a floor. A piecewise quadratic adapted from the one-epoch construction of Cha et al. (2023) rectifies the bridge into a mean displacement of order $\eta\sigma_\star\sqrt{n}\min\{\alpha K, 1\}$. Choosing the larger of these two gadgets and adding one deterministic coordinate gives a matching two-dimensional instance. Curvature rescaling handles large nominal steps without padding or inactive-component arguments.

Contributions. We give the first fixed-step minimax law for the all-inner output of random reshuffling at every finite horizon and every $\eta \le 1/(6L)$. The law answers the lower-bound question in Liu (2026), but also shows why its upper envelope is not pointwise tight. The proof introduces a reset bridge decomposition for the all-inner transient, identifies an exact quadratic cancellation, and transfers the known rectification gadget from epoch endpoints to all-inner averaging.

2 Setup

Let

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \qquad f_i : \mathbb{R}^d \to \mathbb{R},$$

where every component is convex and L-smooth. Let $x^\star$ minimize f and define

$$\sigma_\star^2 = \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x^\star)\|^2.$$

At epoch k, random reshuffling draws an independent uniform permutation π_k and performs

$$x_{k,j+1} = x_{k,j} - \eta \nabla f_{\pi_k(j+1)}(x_{k,j}), \quad j = 0, \dots, n-1, \quad x_{k+1,0} = x_{k,n}. \tag{3}$$

We study the pre-update all-inner average

$$\bar{x}_K = \frac{1}{nK} \sum_{k=0}^{K-1} \sum_{j=0}^{n-1} x_{k,j}. \tag{4}$$

For a constant step, this is exactly the weighted output in Theorem 1 of Liu (2026). It can be maintained online with one accumulator and no additional gradient evaluations. Unlike an epoch-end average, it retains the geometry of the within-epoch path.

For $D, \sigma \geq 0$, define the fixed-step minimax risk

$$\mathfrak{R}_{n,K}(\eta, L, D, \sigma) = \sup \mathbb{E}[f(\bar{x}_K) - f(x^\star)], \tag{5}$$

where the supremum ranges over all dimensions, all finite sums satisfying the assumptions above, and all initial points with $\|x_{0,0} - x^\star\| \leq D$ and $\sigma_\star \leq \sigma$. The lower-bound instance may depend on (η, K), as usual in a minimax statement.

Set

$$\alpha = \eta n L, \qquad \delta = \min\{\alpha, 1\}, \qquad \mathcal{S}_{n,K} = \eta \sigma^2 \delta \left[\frac{1}{K} + \min\{(\delta K)^2, 1\} \right]. \tag{6}$$

For $\alpha \geq 1$, the bracket lies in $[1, 2]$, so $\mathcal{S}_{n,K} \asymp \eta \sigma^2$.

3 The finite-horizon minimax law

Theorem 1 (Fixed-step minimax law)**.** *There are universal constants $c, C > 0$ such that for every $n \geq 8$, $K \geq 1$, $L > 0$, $D, \sigma \geq 0$, and $0 < \eta \leq 1/(6L)$,*

$$\begin{aligned} c \left[\min\left\{ LD^2, \frac{D^2}{\eta n K} \right\} + \mathcal{S}_{n,K} \right] &\leq \mathfrak{R}_{n,K}(\eta, L, D, \sigma) \\ &\leq C \left[\min\left\{ LD^2, \frac{D^2}{\eta n K} \right\} + \mathcal{S}_{n,K} \right]. \end{aligned} \tag{7}$$

The lower bound is attained in dimension at most two. For even n it uses balanced signs. For odd n the quadratic bridge uses one zero sign, while the rectifier uses $n-1$ active components and one zero component.

The theorem separates a deterministic optimization term from a stochastic reshuffling term. It does not assert that every instance follows the three-phase law. In fact, the exact quadratic identity in Section 5 gives an instance whose error vanishes after the bridge phase. The statement is a worst-case characterization at each fixed step and horizon.

The result also explains why the tuned rate in Liu (2026) is correct despite the slack in (1). Minimizing (7) over $0 < \eta \leq 1/(6L)$ gives, up to universal constants,

$$\min\left\{ LD^2, \frac{LD^2}{nK} + \min\left[\frac{\sigma D}{\sqrt{nK}}, \left(\frac{L\sigma^2 D^4}{nK^2} \right)^{1/3} \right] \right\}. \tag{8}$$

The outer cap is necessary when the noise scale is large, since taking $\eta \downarrow 0$ leaves at most the smoothness error LD^2. In the nontrivial regime where the bracketed tuned rate is below this cap, (8) is the rate of Liu (2026). The transient dip concerns fixed steps. Tuning can move the trajectory to a different phase.

4 Upper bound: a reset bridge and one remainder

Translate $x^\star$ to zero and write

$$g_i = \nabla f_i(0), \qquad h_i(x) = \nabla f_i(x) - g_i.$$

Then $\sum_i g_i = 0$ and $\|h_i(x)\| \le L\|x\|$. First run (3) from zero and call this trajectory z. Flatten its $N = nK$ pre-update iterates into $z_0, \dots, z_{N-1}$. Inside each epoch define the reset bridge

$$Y_{k,j} = -\eta \sum_{r=1}^{j} g_{\pi_k(r)}, \qquad j = 0, \dots, n-1. \tag{9}$$

Completed epochs contribute zero to the frozen gradients. Therefore, after flattening $Y_{k,j}$ into Y_t,

$$z_t = Y_t + R_t, \qquad R_t = -\eta \sum_{s<t} h_{i_s}(z_s). \tag{10}$$

The bridge has exact finite-population moments.

Lemma 2 (Reset bridge)**.** *For $j = 0, \dots, n-1$,*

$$\mathbb{E}\|Y_{k,j}\|^2 = \eta^2 \sigma_\star^2 \frac{j(n-j)}{n-1}, \tag{11}$$

$$\mathbb{E}\left\|\frac{1}{N}\sum_{t=0}^{N-1} Y_t\right\|^2 = \frac{\eta^2(n+1)}{12K}\sigma_\star^2. \tag{12}$$

Also,

$$\frac{1}{N}\sum_{t=0}^{N-1} \mathbb{E}\|Y_t\|^2 = \frac{\eta^2(n+1)}{6}\sigma_\star^2. \tag{13}$$

The $1/K$ term in (6) is exactly the variance reduction from averaging the independent centered bridge areas in (12). The second term comes from controlling R.

Lemma 3 (Short-horizon memory)**.** *Let $\theta = \eta L N = \alpha K$. If $\theta \le 1/2$, then*

$$\mathbb{E}\|\bar{z}_K\|^2 \le C_0 \eta^2 n \sigma_\star^2 \left(\frac{1}{K} + \theta^2\right) \tag{14}$$

for a numerical constant C_0.

The proof is a one-line operator estimate. Let P be the strictly lower triangular prefix-sum matrix on N coordinates. With $a_t = \|z_t\|$ and $b_t = \|Y_t\|$, (10) gives $a \le b + \eta L P a$. Since $\|P\|_{2\to 2} \le N$,

$$\|a\|_2 \le 2\|b\|_2, \qquad \|R\|_2 \le 2\theta\|b\|_2.$$

Jensen's inequality, (12), and (13) prove (14). This argument uses no Hessians, commutativity, or dimension restriction.

Smoothness of f now gives

$$\mathbb{E}[f(\bar{z}_K) - f^\star] \le C\eta^2 n L \sigma_\star^2 \left(\frac{1}{K} + (\alpha K)^2\right), \qquad \alpha K \le \tfrac{1}{2}. \tag{15}$$

To restore an arbitrary initialization, couple x and z with the same permutations. Every component gradient map is nonexpansive when $\eta \le 2/L$, so $\|\bar{x}_K - \bar{z}_K\| \le D$. Co-coercivity and smoothness imply

$$f(\bar{x}_K) - f^\star \le 2\big(f(\bar{z}_K) - f^\star\big) + LD^2. \tag{16}$$

When $\alpha K \le 1/2$, the deterministic minimum in (7) equals LD^2, so this proves the desired short-horizon bound.

For $\alpha K > 1/2$, Theorem 1 of Liu (2026) gives

$$\mathbb{E}[f(\bar{x}_K) - f^\star] \le \frac{6D^2}{\eta n K} + 51\eta\sigma_\star^2 \min\{\alpha, 1\}. \tag{17}$$

If $\alpha \le 1$, the bracket in (6) is at least $1/4$ in this region. If $\alpha \ge 1$, it is at least one. The first term in (17) is within a universal factor of the capped deterministic term. This completes the upper bound.

5 LOWER BOUND: CANCELLATION AND RECTIFICATION

The deterministic term follows from identical components $f_i(w) = \lambda w^2/2$, initialized at $w_0 = D$, with $\lambda = \min\{L, (\eta n K)^{-1}\}$. Their geometric all-inner average is a constant fraction of D, which yields

$$f(\bar{w}_K) - f^\star \geq c \min\left\{LD^2, \frac{D^2}{\eta n K}\right\}. \tag{18}$$

It remains to obtain (6). We first expose the fresh bridge with a centered common-Hessian quadratic. Let s_i be balanced signs and set

$$Q_i(u) = \frac{\ell}{2}u^2 + a s_i u, \qquad u_0 = 0. \tag{19}$$

For odd n, one sign is zero. The scale a is adjusted so that $n^{-1}\sum_i a^2 s_i^2 = \sigma^2$. Put $q = 1 - \eta\ell$ and $r = q^n$. Summing all nK update equations and using $\sum_i s_i = 0$ gives the pathwise identity

$$\boxed{\bar{u}_K = -\frac{u_{K,0}}{\eta \ell n K}.} \tag{20}$$

This exact cancellation is special to all-inner averaging.

If

$$V_n(q) = \sum_{j=1}^{n}\left(q^{n-j} - \frac{1}{n}\sum_{r=1}^{n} q^{n-r}\right)^2,$$

finite-population covariance gives

$$\mathbb{E}[Q(\bar{u}_K) - Q(0)] = \frac{\sigma^2 V_n(q)}{2\ell n(n-1)K^2}\,\frac{1-r^{2K}}{1-r^2} \tag{21}$$

for every n, where $Q = n^{-1}\sum_i Q_i = \ell u^2/2$. When $\beta = \eta n \ell$ and βK are small, $V_n(q) \asymp (\eta\ell)^2 n^3$ and the geometric factor is $\asymp K$. Hence

$$\mathbb{E}[Q(\bar{u}_K) - Q(0)] \geq c\frac{\eta^2 n \ell \sigma^2}{K}. \tag{22}$$

For $K \gg 1/\beta$, (21) instead decays as K^{-2}. Common-Hessian quadratics cannot create the floor.

The floor comes from curvature rectification. Consider

$$r_i(v) = \frac{1}{2}\left(\ell \mathbf{1}_{\{v<0\}} + \frac{\ell}{2415}\mathbf{1}_{\{v\geq 0\}}\right)v^2 + a s_i v, \qquad v_0 = 0. \tag{23}$$

Every component is convex and ℓ-smooth. The average objective is minimized at zero and is $\ell/2415$-strongly convex.

Lemma 4 (Finite-horizon rectification)**.** *There are universal constants $b, c_1, c_2 > 0$ such that, whenever $\beta = \eta n \ell \leq b$, random reshuffling on (23) satisfies, with $B = \eta a\sqrt{n}$,*

$$m_{k+1} \geq (1 - c_1\beta)m_k + c_2\beta B, \qquad m_k = \mathbb{E}[v_{k,0}], \quad m_0 = 0. \tag{24}$$

Consequently, if $K \geq \beta^{-2/3}$, then

$$\mathbb{E}[r(\bar{v}_K) - r(0)] \geq c\eta^2 n \ell a^2 \min\{(\beta K)^2, 1\}. \tag{25}$$

For even n the statement uses balanced signs. For odd n, it uses the balanced even construction on $n-1$ active components and one zero component.

The recurrence is the one-epoch rectification estimate proved in Appendix B of Cha et al. (2023). The new all-inner transfer is short. Couple (23) to the common-curvature ℓ quadratic under the same permutation. Monotonicity gives a pathwise lower comparison. Conditioned on an epoch start,

$$\mathbb{E}[v_{k,j} \mid v_{k,0}] \geq (1 - \eta\ell)^j v_{k,0}.$$

Solving (24) gives $m_k \ge cB\min\{\beta k, 1\}$. Averaging over the starts of the first K epochs, and using $K \ge \beta^{-2/3} > 1$, yields $\mathbb{E}\bar{v}_K \ge cB\min\{\beta K, 1\}$. Strong convexity and Jensen's inequality prove (25). Appendix D records the constants and the odd-n reduction.

To cover every nominal step, choose an effective curvature

$$\ell = \frac{\beta}{\eta n}, \qquad \beta = \min\{\alpha, b\}. \tag{26}$$

Then $\ell \le L$. Use (19) when $K \lesssim \beta^{-2/3}$ and (23) otherwise. The two bounds give

$$c\eta\sigma^2\beta\left[\frac{1}{K} + \min\{(\beta K)^2, 1\}\right].$$

Since $\beta/\delta \ge b$ and, for $0 < \beta \le \delta \le 1$,

$$\frac{1}{K} + \min\{(\beta K)^2, 1\} \ge \left(\frac{\beta}{\delta}\right)^2\left[\frac{1}{K} + \min\{(\delta K)^2, 1\}\right],$$

this is a universal constant fraction of (6). Finally, take the orthogonal direct sum of the deterministic coordinate and the selected stochastic coordinate. Block-diagonal smoothness remains at most L, the initial distance is exactly D, and the gradient variance at the minimizer is exactly σ^2. This proves the lower bound in Theorem 1.

6 Relation to Prior Reshuffling Theory

Classical analyses established improved rates for random reshuffling under small-step, large-epoch, strong-convexity, or bounded-gradient conditions (Gürbüzbalaban et al., 2021; Nagaraj et al., 2019; Haochen & Sra, 2019; Ahn et al., 2020; Rajput et al., 2020; Mishchenko et al., 2020; Nguyen et al., 2021). Lower bounds for final iterates and quadratic problems reveal condition-number and epoch transitions (Safran & Shamir, 2020; 2021). Last-iterate and primal-dual analyses have since sharpened output guarantees (Liu & Zhou, 2024; Cai & Diakonikolas, 2025; Cai et al., 2024). Counterexamples to broad reshuffling-versus-replacement comparisons concern noiseless quadratics, matrix inequalities, or different algorithms and outputs (De Sa, 2020).

Several nearby results address different regimes or outputs. The small-epoch analysis of Kim et al. (2025) concerns final-iterate incremental gradient descent under strong convexity. Liu & Zhou (2025) studies nonsmooth last iterates and a one-epoch suffix average. Work on optimized permutations asks when a selected ordering can improve on a random one (Rajput et al., 2022). Recent preprints establish high-probability nonconvex stationarity for random reshuffling (Yu & Li, 2026), analyze block reshuffling and paired reversal (Nguyen et al., 2026), or compare single-shuffle, random-reshuffle, and full-gradient matrix products on quadratics (Peng, 2026). None of these results characterizes the expected smooth-convex risk of the exact all-inner average at a fixed step and every finite horizon.

The closest lower bound is Cha et al. (2023). It treats arbitrary nonnegative combinations of epoch-boundary iterates. Its convex corollary uses a small step and a sufficiently large horizon. We use its one-epoch piecewise-quadratic estimate, but our output contains every inner iterate and our result tracks every finite horizon. The bridge term and the transient dip have no analogue in an endpoint-only average.

The closest upper bound is Liu (2026). It introduced the exact all-inner output in (4), proved (17) for every $\eta \le 1/(6L)$ and every horizon, and explicitly left the corresponding complete lower characterization open. Our theorem closes that direction. It also shows that the stochastic part of (17) is loose by an unbounded factor at $K \asymp \alpha^{-2/3}$ as $\alpha \to 0$.

Constant-step stationary-distribution and bias expansions study a different asymptotic object, usually epoch endpoints under strong monotonicity or additional structure (Emmanouilidis et al., 2026). They are complementary to the finite-horizon convex minimax law here.

7 Discussion

The nonmonotonicity is an output effect, not a claim that running a fixed instance longer must hurt. Minimax instances can depend on the horizon. The law says which mechanism an adversary can activate at each horizon. It also clarifies the role of state memory. Memory creates no fourth minimax scale. During the transient it is contained by the Lipschitz remainder, while in the common-Hessian lower gadget it cancels the bridge exactly.

Our theorem uses the maximum component smoothness L. A sharper law that separates average and maximum smoothness, as in the nonuniform upper theorem of Liu (2026), remains open. The result also concerns the specific all-inner average. Other output rules can have different horizon laws.

Reproducibility Statement

All new arguments are included in the manuscript. The only imported proof ingredient is the published one-epoch rectification estimate of Cha et al. (2023), which Appendix D states with the constants and hypotheses used here. The supplementary verification script checks the exact quadratic identity by permutation enumeration and Monte Carlo simulation, including the pre-update indexing convention. It also enumerates the odd-component reduction and grid-checks the tuned corollary. It exhaustively tests the imported one-epoch inequalities for $n = 8$. No external data or specialized hardware is required.

References

Kwangjun Ahn, Chulhee Yun, and Suvrit Sra. SGD with shuffling: Optimal rates without component convexity and large epoch requirements. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17526–17535, 2020.

Xufeng Cai and Jelena Diakonikolas. Last iterate convergence of incremental methods as a model of forgetting. In *International Conference on Learning Representations*, 2025.

Xufeng Cai, Cheuk Yin Lin, and Jelena Diakonikolas. Tighter convergence bounds for shuffled SGD via primal-dual perspective. In *Advances in Neural Information Processing Systems*, volume 37, pp. 72475–72524, 2024.

Jaeyoung Cha, Jaewook Lee, and Chulhee Yun. Tighter lower bounds for shuffling SGD: Random permutations and beyond. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3855–3912. PMLR, 2023.

Christopher M. De Sa. Random reshuffling is not always better. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5957–5967, 2020.

Konstantinos Emmanouilidis, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and René Vidal. Shuffling the data, stretching the step-size: Sharper bias in constant step-size SGD. In *International Conference on Learning Representations*, 2026.

Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186:49–84, 2021.

Jeff Haochen and Suvrit Sra. Random shuffling beats SGD after finite epochs. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2624–2633. PMLR, 2019.

Yujun Kim, Jaeyoung Cha, and Chulhee Yun. Incremental gradient descent with small epoch counts is surprisingly slow on ill-conditioned problems. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29913–29968. PMLR, 2025.

Zijian Liu. Random reshuffling dominates stochastic gradient descent. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 4859–4882. PMLR, 2026.

Zijian Liu and Zhengyuan Zhou. On the last-iterate convergence of shuffling gradient methods. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32471–32508. PMLR, 2024.

Zijian Liu and Zhengyuan Zhou. Improved last-iterate convergence of shuffling gradient methods for nonsmooth convex optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40152–40193. PMLR, 2025.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17309–17320, 2020.

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. SGD without replacement: Sharper rates for general smooth convex functions. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4703–4711. PMLR, 2019.

Lam M. Nguyen, Quoc Tran-Dinh, Dzung T. Phan, Phuong Ha Nguyen, and Marten van Dijk. A unified convergence analysis for shuffling-type gradient methods. *Journal of Machine Learning Research*, 22(207):1–44, 2021.

Lam M. Nguyen, Dzung T. Phan, and Jayant Kalagnanam. Learning to shuffle: Block reshuffling and reversal schemes for stochastic optimization. *arXiv preprint arXiv:2604.00260*, 2026.

Binghui Peng. A resolution of the SS–RS–GD inequalities. *arXiv preprint arXiv:2607.22620*, 2026.

Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of SGD without replacement. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7964–7973. PMLR, 2020.

Shashank Rajput, Kangwook Lee, and Dimitris Papailiopoulos. Permutation-based SGD: Is random optimal? In *International Conference on Learning Representations*, 2022.

Itay Safran and Ohad Shamir. How good is SGD with random shuffling? In *Proceedings of the Thirty-Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3250–3284. PMLR, 2020.

Itay Safran and Ohad Shamir. Random shuffling beats SGD only after many epochs on ill-conditioned problems. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15151–15161, 2021.

Hengxu Yu and Xiao Li. High probability guarantees for random reshuffling. *arXiv preprint arXiv:2311.11841*, 2026.

A FINITE-POPULATION IDENTITIES

We prove Lemma 2. Let $G_1, \dots, G_n \in \mathbb{R}^d$ satisfy $\sum_i G_i = 0$ and $n^{-1} \sum_i \|G_i\|^2 = \sigma_G^2$. If π is uniform, then

$$\mathbb{E}\langle G_{\pi(r)}, G_{\pi(s)} \rangle = \begin{cases} \sigma_G^2, & r = s, \\ -\sigma_G^2/(n-1), & r \neq s. \end{cases}$$

It follows directly that

$$\mathbb{E} \left\| \sum_{r=1}^{j} G_{\pi(r)} \right\|^2 = \sigma_G^2 \frac{j(n-j)}{n-1}. \tag{27}$$

Taking $G_i = g_i$ proves (11). Summing (27) over $j = 0, \dots, n-1$ and using

$$\sum_{j=0}^{n-1} j(n-j) = \frac{n(n-1)(n+1)}{6}$$

proves (13).

For the area of one epoch,

$$A_\pi = \frac{1}{n}\sum_{j=0}^{n-1} Y_j = -\frac{\eta}{n}\sum_{r=1}^{n-1}(n-r)g_{\pi(r)}.$$

For arbitrary scalar weights w_r, the covariance above gives

$$\mathbb{E}\left\|\sum_{r=1}^{n} w_r g_{\pi(r)}\right\|^2 = \frac{n\sigma_\star^2}{n-1}\sum_{r=1}^{n}(w_r-\bar{w})^2. \tag{28}$$

With $w_r = n-r$ and $\sum_r (w_r-\bar{w})^2 = n(n^2-1)/12$, we obtain

$$\mathbb{E}\|A_\pi\|^2 = \frac{\eta^2(n+1)}{12}\sigma_\star^2.$$

The epoch areas are independent and centered. Their average has variance $1/K$ times this quantity, proving (12).

B Complete upper-bound proof

We first prove Lemma 3. Index the $N = nK$ pre-update iterates from zero. Let $P \in \mathbb{R}^{N\times N}$ have entries $P_{ts} = 1$ if $s < t$ and zero otherwise. Equation (10) and $\|h_i(z)\| \le L\|z\|$ imply componentwise

$$a \le b + \eta L P a, \qquad a_t = \|z_t\|, \quad b_t = \|Y_t\|.$$

Since $\|P\|_{2\to 2} \le \|P\|_{\mathrm{F}} < N$ and $\theta = \eta L N \le 1/2$,

$$\|a\|_2 \le \|b\|_2 + \theta\|a\|_2 \le 2\|b\|_2. \tag{29}$$

Likewise, the sequence of remainder norms satisfies

$$\|R\|_2 \le \eta L\|Pa\|_2 \le 2\theta\|b\|_2. \tag{30}$$

Jensen's inequality gives

$$\mathbb{E}\|\bar{R}\|^2 \le \frac{1}{N}\mathbb{E}\|R\|_2^2 \le 4\theta^2\frac{1}{N}\mathbb{E}\|b\|_2^2 = \frac{2(n+1)}{3}\eta^2\theta^2\sigma_\star^2.$$

Together with (12) and $\|u+v\|^2 \le 2\|u\|^2 + 2\|v\|^2$, this proves (14) with, for example, $C_0 = 4$ for $n \ge 2$.

For completeness, we justify the initialization coupling. If $T_i(x) = x - \eta\nabla f_i(x)$, co-coercivity gives

$$\|T_i(x) - T_i(y)\|^2 \le \|x-y\|^2 - (2\eta/L - \eta^2)\|\nabla f_i(x) - \nabla f_i(y)\|^2.$$

Thus T_i is nonexpansive for $\eta \le 2/L$. Coupled trajectories remain within D, and so do their averages. For any u, v, smoothness gives

$$f(u) - f^\star \le f(v) - f^\star + \langle \nabla f(v), u - v\rangle + \frac{L}{2}\|u-v\|^2.$$

The standard inequality $\|\nabla f(v)\|^2 \le 2L(f(v) - f^\star)$ and Young's inequality yield (16).

If $\theta \le 1/2$, then $D^2/(\eta n K) = LD^2/\theta \ge 2LD^2$, so the capped deterministic term is LD^2. Equations (15) and (16) give the upper half of (7) in this region.

If $\theta > 1/2$, apply (17). When $\theta \in (1/2, 1)$, $D^2/(\eta n K) < 2LD^2$, and when $\theta \ge 1$ it is the smaller deterministic term. Hence the first term in (17) is at most twelve times the capped term. If $\alpha \le 1$, then

$$K^{-1} + \min\{(\alpha K)^2, 1\} \ge 1/4.$$

If $\alpha \ge 1$, then $\delta = 1$ and the same bracket is at least one. This absorbs the stochastic term in (17) and completes the proof.

C EXACT COMMON-HESSIAN CANCELLATION

We derive (21) and its lower consequence. First assume n is even and $s_i \in \{-1, 1\}$ are balanced. For one epoch,

$$u_{k+1,0} = ru_{k,0} + \xi_k, \qquad \xi_k = -\eta a \sum_{j=1}^{n} q^{n-j} s_{\pi_k(j)}.$$

The innovations are independent and centered. Equation (28) gives

$$\mathbb{E}\xi_k^2 = \frac{\eta^2 n a^2}{n-1} V_n(q).$$

Starting from zero,

$$\mathbb{E}u_{K,0}^2 = \frac{\eta^2 n a^2}{n-1} V_n(q) \sum_{k=0}^{K-1} r^{2k}.$$

The sum of the nK scalar update equations is

$$u_{K,0} - u_{0,0} = -\eta\ell \sum_{k,j} u_{k,j} - \eta a \sum_{k,j} s_{\pi_k(j)}.$$

The last term vanishes epoch by epoch. This proves (20). Substitution into $Q(\bar{u}_K) - Q(0) = \ell \bar{u}_K^2/2$ proves (21).

We record explicit bounds. If $\beta = \eta n \ell \leq 1/4$, then $q^{n-1} \geq 1 - \beta \geq 3/4$. The pairwise variance identity gives

$$\begin{aligned} V_n(q) &= \frac{1}{2n} \sum_{p,r=0}^{n-1} (q^p - q^r)^2 \\ &\geq \frac{q^{2n-2}(\eta\ell)^2}{2n} \sum_{p,r=0}^{n-1} (p-r)^2 \geq c(\eta\ell)^2 n^3. \end{aligned}$$

If also $\beta K \leq 1/4$, then $r^{2k} \geq (1-\beta)^{2K} \geq c$ for $k < K$, so $\sum_{k<K} r^{2k} \geq cK$. Equation (21) now proves (22).

For odd n, choose $(n-1)/2$ signs of each type and one zero. Put $\rho_n = n^{-1} \sum_i s_i^2 = (n-1)/n$ and $a = \sigma/\sqrt{\rho_n}$. The weighted covariance formula has $a^2 \rho_n = \sigma^2$ in place of a^2, so the same calculation and constants apply.

D THE RECTIFICATION LEMMA

We give the promised reduction to the proved one-epoch estimates of Cha et al. (2023). Their Appendix B considers exactly (23) for even n, with a denoted by ν.

Lemma 5 (Published one-epoch estimates)**.** *Let $n \geq 8$ be even and $\beta = \eta n \ell \leq 1/161$. For an epoch starting from a fixed v, let v^+ denote its endpoint. Lemmas B.2–B.4 and the calculation following them in Cha et al. (2023) give, for $v \geq 0$,*

$$\mathbb{E}[v^+ \mid v] \geq (1 - 3\beta/4)v + \frac{\eta^2 \ell a n^{3/2}}{18000}, \tag{31}$$

and, for $v < 0$,

$$\mathbb{E}[v^+ \mid v] \geq (1 - 160\beta/161)v. \tag{32}$$

If the first epoch starts nonnegative, every later epoch start is nonnegative with probability at least 1/2.

Since $v < 0$ reverses the comparison between the two contraction coefficients, conditioning on the sign of the start gives

$$m_{k+1} \geq (1 - 3\beta/4)m_k + \frac{\beta B}{36000}, \qquad B = \eta a\sqrt{n}. \tag{33}$$

This is (24). Taking a smaller universal b if needed, iteration from $m_0 = 0$ yields

$$m_k \geq cB \min\{\beta k, 1\}. \tag{34}$$

We next transfer endpoint drift to the all-inner output. Let $\widetilde{v}_j$ be the trajectory with the same linear signs and common curvature ℓ. For $\eta\ell \leq 1$, both component update maps are nondecreasing and

$$T_i^{\mathrm{rect}}(x) \geq T_i^{\mathrm{quad}}(x) \quad \text{for every } x.$$

Induction gives $v_j \geq \widetilde{v}_j$ pathwise. Conditional on the epoch start, uniform permutation marginals and balanced signs give

$$\mathbb{E}[\widetilde{v}_j \mid v_0] = (1 - \eta\ell)^j v_0.$$

Therefore

$$\mathbb{E}v_{k,j} \geq (1 - \eta\ell)^j m_k \geq (1 - \beta) m_k.$$

If $K \geq \beta^{-2/3}$ and b is small, then $K \geq 2$. Averaging (34) over $k = 0, \ldots, K-1$ gives

$$\mathbb{E}\bar{v}_K \geq cB \min\{\beta K, 1\}. \tag{35}$$

The average objective in (23) is $\ell/2415$-strongly convex and minimized at zero. Jensen's inequality gives

$$\mathbb{E}[r(\bar{v}_K) - r(0)] \geq \frac{\ell}{4830} (\mathbb{E}\bar{v}_K)^2,$$

which proves (25).

For odd n, let $m = n - 1$, use the even construction on m active components, and set the final component identically to zero. Choose $a = \sigma\sqrt{n/m}$ on the active components. Then $n^{-1}\sum_i |r_i'(0)|^2 = \sigma^2$. A uniform permutation of all n components induces a uniform relative order of the active components and an independent uniform insertion gap $J \in \{0, \ldots, m\}$ for the zero component. Its update is the identity, so the epoch endpoint is exactly the endpoint of the m-component active trajectory. The cited even-instance recurrence therefore applies with

$$\beta_m = \eta m\ell = \frac{m}{n}\beta, \qquad B_m = \eta a\sqrt{m} = \eta\sigma\sqrt{n} = \sqrt{\frac{m}{n}}\, B.$$

Because $m/n \geq 8/9$, replacing β_m by β changes only universal constants in (24) and (34).

It remains to check the all-inner average. Write $y_0, \ldots, y_m$ for the active trajectory within one epoch. The n pre-update states have sum

$$\sum_{j=0}^{m-1} y_j + y_J.$$

The same pathwise quadratic comparison used above gives $\mathbb{E}y_j \geq (1 - \eta\ell)^j m_k \geq 0$ for every $j \in \{0, \ldots, m\}$. Thus the inserted duplicate has nonnegative expectation, and the full all-inner mean is at least m/n times the active all-inner mean. Finally, the average objective is m/n times the active average objective, hence it is $(m/n)\ell/2415$-strongly convex. Applying Jensen's inequality and using $m/n \geq 8/9$ proves (25) with another universal constant.

E Completing the Lower Bound

Choose a universal $b > 0$ small enough that $b \leq 1/161$ and $b^{1/3} \leq 1/4$. Let $\beta = \min\{\alpha, b\}$ and choose ℓ as in (26). Then $\ell \leq L$ and $\eta n\ell = \beta$.

If $K \leq \beta^{-2/3}$, then $\beta K \leq \beta^{1/3} \leq 1/4$ and Appendix C gives

$$\mathbb{E}[Q(\bar{u}_K) - Q(0)] \geq c\eta^2 n\ell\sigma^2/K.$$

In this regime, $(\beta K)^2 \leq 1/K$, so this is a constant fraction of

$$\eta^2 n\ell\sigma^2 \left[K^{-1} + \min\{(\beta K)^2, 1\}\right].$$

If $K \geq \beta^{-2/3}$, Lemma 4 applies and $K^{-1} \leq (\beta K)^2$. Its lower bound is again a constant fraction of the same display. Thus one of the two one-dimensional stochastic gadgets always supplies

$$c\eta\sigma^2\beta \left[K^{-1} + \min\{(\beta K)^2, 1\}\right]. \tag{36}$$

Let $\delta = \min\{\alpha, 1\}$. We have $\beta/\delta \geq b$. For $r = \beta/\delta \leq 1$,

$$\min\{(\beta K)^2, 1\} \geq r^2 \min\{(\delta K)^2, 1\}, \qquad K^{-1} \geq r^2 K^{-1}.$$

Hence (36) is at least $cb^3 \mathcal{S}_{n,K}$.

Finally, verify (18). Put $T = nK$ and $\lambda = \min\{L, (\eta T)^{-1}\}$. The pre-update iterates are $w_t = (1 - \eta\lambda)^t D$, and

$$\bar{w}_K = D \frac{1 - (1 - \eta\lambda)^T}{T\eta\lambda}.$$

If $\eta LT \leq 1$, then $\lambda = L$ and $\bar{w}_K \geq cD$. If $\eta LT \geq 1$, then $\lambda = (\eta T)^{-1}$ and again $\bar{w}_K \geq cD$. Multiplying by $\lambda/2$ proves (18). The orthogonal direct sum described in the main text finishes the proof of Theorem 1.

F Optimizing the step

We briefly derive (8). Sending η to zero gives the initial smoothness cap LD^2. Any step that improves on this cap has $\eta LnK \gtrsim 1$, so its deterministic contribution is $D^2/(\eta nK)$. In the regime $\alpha \geq 1$, balancing this term with $\eta\sigma^2$ gives $\sigma D/\sqrt{nK}$. In the small-step regime after rectification has saturated, balancing it with $\eta^2 nL\sigma^2$ gives $(L\sigma^2 D^4/(nK^2))^{1/3}$. A step of order $1/L$ also contributes the baseline $LD^2/(nK)$. Taking the better admissible balance and then the initial-error cap yields (8). The lower half of Theorem 1 gives the matching capped expression.